# A high-dimensional two-sample test for the mean using random subspaces

Måns Thulin

Department of Mathematics, Uppsala University

**Abstract**

A common problem in genetics is that of testing whether a set of highly dependent gene expressions differ between two populations, typically in a high-dimensional setting where the data dimension is larger than the sample size. Most high-dimensional tests for the equality of two mean vectors rely on naive diagonal or trace estimators of the covariance matrix, ignoring dependencies between variables. A test recently proposed by Lopes et al. (2012) implicitly incorporates dependencies by using random pseudo-projections to a lower-dimensional space. Their test offers higher power when the variables are dependent, but lacks desirable invariance properties and relies on asymptotic p-values that are too conservative. We illustrate how a permutation approach can be used to obtain p-values for the Lopes et al. test and how modifying the test using random subspaces leads to a test statistic that is invariant under linear transformations of the marginal distributions. The resulting test does not rely on assumptions about normality or the structure of the covariance matrix. We show by simulation that the new test has higher power than competing tests in realistic settings motivated by microarray gene expression data. We also discuss the computational aspects of high-dimensional permutation tests and provide an efficient R implementation of the proposed test.

**Keywords:** computational statistics; gene expression data; gene-set testing; high-dimensional data; large $p$ small $n$; permutation test; random subspace; test about the mean; two-sample problem.

## 1 Introduction

A commonly encountered problem in modern genetic research, geological imaging, signal processing, astrometry and finance is that of comparing the mean vectors

of two populations. In many of today's applications, the data averts analysis by classic statistical methods as the data dimension $p$ typically is larger than the sample size $n$, which causes most standard procedures to break down. When $p < n$, comparisons of this type are usually done using Hotelling's $T^2$ test. In the high-dimensional setting where $p \geq n$, the sample covariance matrix is not invertible, meaning that Hotelling's test no longer can be used.

In this paper we discuss two-sample tests in a high-dimensional setting where the variables have a non-negligible dependence structure. While the tests are discussed in the context of gene-set testing, we stress that they are equally applicable to other fields in which high-dimensional data occur.

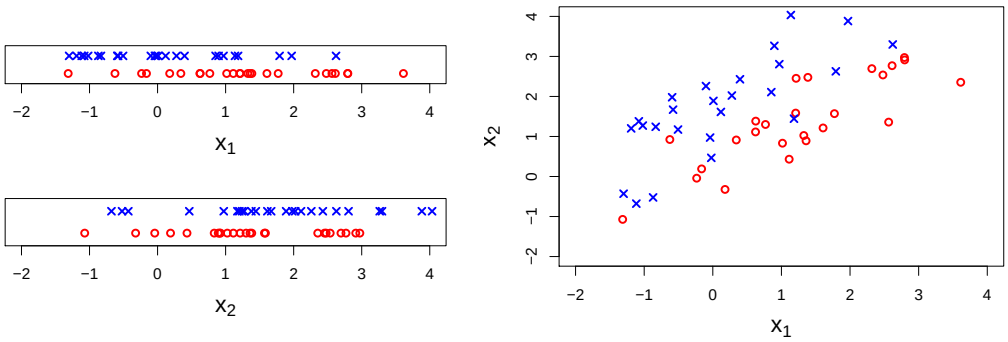
In genetic research, one is often interested in identifying differentially expressed genes between two groups of patients based on data from a microarray experiment. Genes, however, do not function in isolation. Rather, they work together in complex networks. It is therefore often of greater interest to search for sets of genes, rather than individual genes, that are differentially expressed (Barry et al., 2005; Efron & Tibshirani, 2007; Goeman & Bühlmann, 2007; Newton et al., 2007; Nettleton et al., 2008; Barry et al., 2008; Chen & Qin, 2010). These gene-sets are determined a priori, typically by utilizing databases such as Gene Ontology[1] or Kyoto Encyclopedia of Genes and Genomes[2] or by grouping genes with similar chromosomal locations together.

A common approach to finding differentially expressed gene-sets is to use a two-step procedure, starting by performing individual tests for each gene. These gene-level tests are then aggregated into a single test for the entire gene set. Much, if not all, of the multivariate structure of the data set is lost when gene-level test are used. Goeman & Bühlmann (2007), Efron (2007) and Gatti et al. (2010) demonstrated several problems with common tests based on this approach, including very high rates of false positives. The empirical Bayes methods applied e.g. by Efron et al. (2001) suffer from similar issues, caused by correlations between gene expressions (Qui et al., 2005).

By using a truly multivariate test for the gene set, it is possible not only to take the multivariate dependence structure of the gene expressions into account, but to gain more power from these dependencies, as illustrated in Figure 1. There are three approaches to modifying the covariance estimator in Hotelling's test statistic to allow for high-dimensional inference. The first approach is to use prior information about the covariance structure to estimate the covariance matrix. Jacob et al. (2012) presented such a test in the setting where location shift between the two populations is related to a known graph structure describing the dependence between the genes.

[1] http://www.geneontology.org/
[2] http://www.genome.jp/kegg/

Figure 1: Comparison of two bivariate data sets. The red circles have population mean vector $(1, 1)$ whereas the blue crosses have population mean vector $(0.25, 1.75)$. The shifts are difficult to detect by looking at the marginal distributions, but become evident when comparing the joint distributions.




The second approach is to assume that the covariance matrix has a simple diagonal structure. The tests proposed by Bai & Saranadasa (1996), Srivastava (2007), Srivastava & Du (2008), Chen & Qin (2010), Srivastava et al. (2013) rely on imposing this particular structure on the covariance matrix, assuming the expressions of different genes to be independent. This is an unrealistic assumption for gene expressions, where genetic regulatory networks tend to cause the expression to be highly correlated. The assumption is equally unrealistic in many other biomedical problems. As a further example, the prevalence of an allele is typically highly correlated with the prevalence of other alleles on neighbouring loci.

The third approach is to use an estimator that allows for dependence, but that can be used in the absence of prior information. Recently, Lopes et al. (2012) proposed a test in which the data is randomly pseudo-projected into several lower-dimensional spaces. Hotelling's $T^2$ statistic is computed for each pseudo-projection, and the result is then averaged over all pseudo-projections. Lopes et al. showed by asymptotic arguments and a simulation study that their test has substantially higher power than competing tests when the variables are correlated. There are however two downsides to their proposed method. First, it relies on an asymptotic null distribution derived under the assumption of normality. For finite sample sizes, this often leads to far too conservative p-values. Second, the test statistic is not invariant under linear transformations of the marginal distributions. This is a serious drawback, as it is common for genetic data to be rescaled by dividing the marginal distributions by their respective standard deviations.

In this paper, we show how accurate p-values for the Lopes et al. test can be obtained by using random permutations. We then propose a modified test statistic,

which uses random subspaces instead of random pseudo-projections. The new test statistic is, conditioned on the random subspaces chosen, invariant under linear transformations of the marginal distributions.

A common problem in gene-set testing is that of identifying gene-sets, or pathways, that are related to cancer. For a pathway to induce cancer, a mutation must have occurred in at least one of its genes. Depending on where in this pathway the gene is located, the mutation can cause changes in the expressions of only a handful of genes or in all genes in the pathway. Motivated by this problem setting, we perform a Monte Carlo comparison of four multivariate two-sample tests and two tests based on gene-level $t$-tests. We compare the type I error rates and powers of the tests under different models for pathway dependencies and mutation locations. Some tests that require normality are modified so that p-values are computed using permutations rather than asymptotic null distributions, resulting in better type I error rates as well as higher power. We also contrast the invariance properties of the test statistics. While invariance properties tend to be overlooked in the biomedical literature, they are of great importance in multivariate testing and need to be taken into account when choosing which test to use.

High-dimensional permutation tests are heavily computer-intensive. For that reason, we discuss some computational aspects of such tests, and show how to efficiently implement the proposed test in R.

Methods based on random projections and random subspaces have not been studied to a great extent in the statistical literature, but are common in machine learning, where these techniques mainly have been been used for clustering and classification. See Durand & Atkison (2011) and Bingham & Mannila (2001) for reviews and some applications and Varmuza et al. (2010) for applications in chemometrics. Most authors have used only a single random projection or subspace, although there are a few exceptions, including the the recent paper by Lopes et al. (2012). Cuesta-Albertos et al. (2006) used multiple random projections for goodness-of-fit testing but did not find the increase in power to be large enough to motivate the added computational complexity. Fern & Brodley (2003) used multiple random projections for clustering, in a manner that bears resemblance to the algorithms presented in this paper, and found that it improved the performance of their clustering algorithms. Recently, Henrion et al. (2011) proposed a subspace method for outlier detection in high-dimensional data sets that is somewhat similar to the random subspaces test presented in the present paper. Their method differs from ours in several ways, the most important difference being that their goal is to give an anomaly score *to each observation*, rather than to compute a statistic that can be used for inference about the underlying population. Also worth mentioning is a recent paper by Wei et al. (2013), who described a general hypothesis testing framework based on a non-random projection to the real line,

determined by a linear classifier.

The rest of the paper is organised as follows. In Section 2 we review the Lopes et al. test and discuss its drawbacks. In Section 3 we propose a new test based on random subspaces. In Section 4 we compare several gene-set tests in terms of invariance, type I error rates and power. In Section 5 we discuss the computational aspects of the new test. The text concludes with a discussion in Section 6 and an appendix with implementations and examples in R.

# 2 The Lopes et al. test

## 2.1 Setting

We consider two random samples of size $n_X$ and $n_Y$ from independent $p$-dimensional random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, with mean vectors $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$ and covariance matrices $\boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma}$. We assume that $n_X + n_Y - 2 \geq p$. Our aim is to test whether $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$.

One could argue that when testing the hypothesis $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ for two groups of patients, say a control group and a group of cancer patients, we should not expect the covariance matrices of the two groups to coincide. Srivastava et al. (2013) proposed a solution to this high-dimensional Behrens–Fisher problem. The hypothesis that we wish to test using mean vectors is however often not strictly speaking $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$, but rather that the two multivariate distributions agree: $\boldsymbol{F}_X = \boldsymbol{F}_Y$. If the genes under consideration have no connection to cancer, there is no reason to expect the covariance matrix for their expressions to be any different from that of the control group. Under the null hypothesis that the two distributions are equal we should therefore assume that $\boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_Y$.

## 2.2 The test statistic

Lopes et al. (2012) proposed a two-sample test of $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ that uses random projections to $k$-dimensional subspaces. Formally, assume that samples $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_X})$ and $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_Y})$ are given. For $k \leq n_X + n_Y - 2$, let $\boldsymbol{P}_k$ be a $k \times p$ random matrix with i.i.d. $N(0, 1)$ elements, independent of the data.

The Lopes et al. statistic is computed by averaging the Hotelling's $T^2$ statistics over several random projections $\boldsymbol{P}_k'\boldsymbol{X}_1, \ldots, \boldsymbol{P}_k'\boldsymbol{Y}_{n_Y}$. The algorithm is as follows.

**Algorithm 1: The Lopes et al. statistic $T_L$**

1. Generate a $k \times p$ random matrix $\boldsymbol{P}_k$ with i.i.d. $N(0,1)$ elements.
2. Calculate the test statistic $T_i^2$ based on the $n_X$ and $n_Y$ observations of the vectors $\boldsymbol{X}_i = \boldsymbol{P}_k'\boldsymbol{X}$ and $\boldsymbol{Y}_i = \boldsymbol{P}_k'\boldsymbol{Y}$.
3. Repeat steps 1-2 $B_1$ times, obtaining the test statistics $T_1^2, \ldots, T_{B_1}^2$.
4. Obtain the resulting test statistic as the function $T_L = B_1^{-1} \sum_{i=1}^{B_1} T_i^2$.

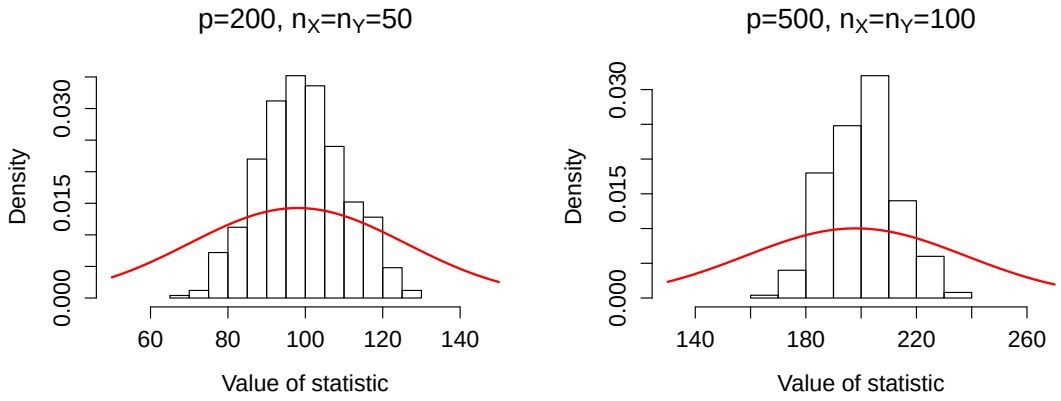
It should be noted that "projection" is used in a loose sense here, since $\boldsymbol{P}_k$ usually isn't a projection matrix. Lopes et al. derived the asymptotic null distribution and power function of their test under normality and derived conditions under which the test is asymptotically more powerful than the tests of Chen & Qin (2010) and Srivastava & Du (2008).

In the following sections, we outline two drawbacks to this test and how the test procedure can be modified to account for these.

## 2.3 The null distribution

The first drawback is that the asymptotic null distribution of the test statistic is a poor approximation of the null distribution for small sample sizes, such as the $p = 200$, $n_X = n_Y = 50$ setting studied in Section 4. Two examples of this phenomenon are given in Figure 2. The p-values obtained from the asymptotic null distribution are highly conservative. For this reason, we instead propose using the permutation distribution of the data to compute the p-values. Let $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{n_X+n_Y}) = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_X}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_Y})$ denote the entire sample. The number of permutations is usually too large for it to be computationally feasible to use the exact permutation distribution of $\boldsymbol{Z}$. We can however obtain approximate p-values by using random permutations to approximate the permutation distribution. The algorithm is as follows.

Figure 2: Histograms of $T_L$ for 500 samples simulated under the null hypothesis. The asymptotic null distribution is shown in red.




**Algorithm 2: A random permutation test**

1. Given a test statistic $T$, compute $T_{obs} = T(\boldsymbol{X}, \boldsymbol{Y})$.
2. Draw $n_X$ integers $i_1, \ldots, i_{n_X}$ from $1, 2, \ldots, n_X + n_Y$ without replacement. Let $j_1, \ldots, j_{n_Y}$ denote the numbers that were not chosen.
3. Calculate the test statistic $T$ for $\boldsymbol{X}^* = (\boldsymbol{Z}_{i_1}, \boldsymbol{Z}_{i_2}, \ldots, \boldsymbol{Z}_{i_{n_X}})$ and $\boldsymbol{Y}^* = (\boldsymbol{Z}_{j_1}, \boldsymbol{Z}_{j_2}, \ldots, \boldsymbol{Z}_{j_{n_Y}})$.
4. Repeat steps 1-2 $B_2$ times, obtaining the test statistics $T_1, \ldots, T_{B_2}$.
5. Compute the p-value of the test as $B_2^{-1} \sum_{i=1}^{B_2} \mathbb{I}(T_i \geq T_{obs})$.

## 2.4 Invariance properties

The second drawback is that the statistic $T_L$ lacks desirable invariance properties. Ideally, the result of a statistical test should not depend on the location and scale on which the measurements have been obtained. Unfortunately, when $p > n_X + n_Y - 2$, no two-sample test statistic that is a function of the sample can be affine invariant (Lehmann, 1959, p. 318), i.e. invariant under all linear transformations. It is however possible for a test statistic to be invariant under a smaller group of linear transformations.

In order to facilitate interpretability, gene expression data is often rescaled so that all variables have standard deviation 1, meaning that $\boldsymbol{\Sigma}$ is a correlation matrix. This rescaling is a linear transformation of the marginal distributions,

$\boldsymbol{X} \to \boldsymbol{DX}$, with $\boldsymbol{D}$ being a diagonal matrix with diagonal elements $1/\sigma_i$, where $\sigma_i$ is the $i$:th standard deviation. In practice these standard deviations are almost invariably estimated. Invariance under such linear transformations of the marginal distributions is arguably the most important invariance property for a high-dimensional gene-set test.

$T_L$ is not invariant under this kind of linear transformations, even if we condition on the random projections $\boldsymbol{P}_k$. Some R code that illustrates this is given in Appendix B. Using the data given in the appendix, for which $p = 20$ and $n_X = n_Y = 5$, we obtained the p-value 0.381 when computing $T_L$ for the raw data with $k = 4$ and $B_1 = B_2 = 1000$. We then standardized the data by dividing by the marginal sample standard deviations. Keeping the random projections and permutations fixed, the p-value was 0.003 after standardization. For the Lopes et al. test, standardization can turn a non-significant gene-set into a significant one.

In Section 3 we propose a modification of the Lopes et al. test that leads to a test statistic that is invariant under linear transformations of the marginal distributions. The invariance properties of several two-sample test statistics are compared in Section 4.2.

# 3 A random subspaces test

## 3.1 Motivation and test procedure

Now, let $\boldsymbol{X}_{(i)}$ and $\boldsymbol{Y}_{(i)}$, $i = 1, 2, \ldots, \binom{p}{k}$, denote the $k$-dimensional subvectors of $\boldsymbol{X}$ and $\boldsymbol{Y}$, with $k \leq n_X + n_Y - 2$. If $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ then $\mathrm{E}(\boldsymbol{X}_{(i)}) = \mathrm{E}(\boldsymbol{Y}_{(i)})$ for all $i$. Tests of the hypothesis $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ can therefore be based on tests of the hypotheses $\mathrm{E}(\boldsymbol{X}_{(i)}) = \mathrm{E}(\boldsymbol{Y}_{(i)})$.

$\binom{p}{k}$ is usually extremely large in the high-dimensional setting and studying all $k$-dimensional subvectors of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is not feasible. As with the Lopes et al. test, we can however randomly select $B_1$ $k$-dimensional subvectors and compute $T^2$ for each subvector, after which a conclusion about the $p$-dimensional hypothesis can be drawn by averaging the test statistics for the different subvectors. Random subspace methods have previously been successfully applied to machine learning problems, e.g. by Bertoni et al. (2005) and Lai et al. (2006).

The heuristic motivation for this procedure is that if the deviations from the null hypothesis are spread out over many variables, they will likely be detected in many subvectors. On the other hand, if the deviations are concentrated in just a few variables, those variables are likely to be investigated if $B$ is large enough.

The algorithm for computing the random subspaces test statistic, denoted $T_{rs}$, is described below. In the rest of the paper the test statistic $T_i^2$ will be Hotelling's statistic, but the test can also be carried out using other statistics.

**Algorithm 3: Computing the random subspace statistic $T_{rs}$**

1. Draw $k \leq n_X + n_Y - 2$ integers $i_1, \ldots, i_k$ from $1, 2, \ldots, p$ without replacement.
2. Calculate the test statistic $T_i^2$ based on the $n_X$ and $n_Y$ observations of the subvectors $\boldsymbol{X}^*_{(i)}$ and $\boldsymbol{Y}^*_{(i)}$, where the subvector with index $(i)$ is in the subspace consisting of the dimensions $i_1, \ldots, i_k$.
3. Repeat steps 1-2 $B_1$ times, obtaining the test statistics $T_1^2, \ldots, T_{B_1}^2$.
4. Obtain the resulting test statistic as the function $T_{rs} = B_1^{-1} \sum_{i=1}^{B_1} T_i^2$.

To compute the p-value for the test, we propose using the permutation distribution, i.e. using Algorithm 2 with the statistic $T_{rs}$.

## 3.2 Properties and relation to the Lopes et al. test

While Lopes et al. proposed using $\boldsymbol{P}_k$ with i.i.d. $N(0,1)$ elements, in their theoretical investigations they studied a more general test procedure, where $\boldsymbol{P}_k$ can be generated by some other distribution.

**Proposition 1.** *$T_{rs}$ is a special case of the general random projections test studied in Lopes et al. (2012).*

To see this, let $\mathbf{1}_i = (j_1, j_2, \ldots, j_p)$ be a $p$-vector with $j_h = \mathbb{I}_i(h)$. If $i_1, \ldots, i_k$ are drawn uniformly at random from $\{1, 2, \ldots, p\}$ without replacement and if $\boldsymbol{P}_k = (\mathbf{1}_{i_1}, \mathbf{1}_{i_2}, \ldots, \mathbf{1}_{i_k})$ then the projected sample is simply the part of the sample that resides in the $k$-dimensional subspace given by the indices $i_1, \ldots, i_k$. Thus the random subspaces test and the general random projections test coincide for this particular choice of $\boldsymbol{P}_k$.

The asymptotic results of Lopes et al. (2012) hold for any random matrices $\boldsymbol{P}_k$ that have full rank with probability 1. Since the $\boldsymbol{P}_k$ described above always has full rank, we have the following result.

**Proposition 2.** *$T_{rs}$ and $T_L$ are asymptotically equivalent.*

Consequently, the asymptotic relative efficiency of the random subspaces test with respect to the Chen–Qin test is given in Theorem 3 of Lopes et al. (2012) and the asymptotic relative efficiency with respect to the Srivastava–Du test is given in Theorem 4 of said paper.

While asymptotically equivalent, $T_L$ and $T_{rs}$ differ in their finite-sample properties, in particular regarding invariance:

**Proposition 3.** *Let $\boldsymbol{D}$ be a diagonal $p \times p$ real matrix with nonzero diagonal elements and let $\boldsymbol{d} \in \mathbb{R}^p$. Conditioned on the random subspaces chosen, $T_{rs}$ is invariant under the linear transformations $(\boldsymbol{X}, \boldsymbol{Y}) \to (\boldsymbol{DX} + \boldsymbol{d}, \boldsymbol{DY} + \boldsymbol{d})$.*

Unlike e.g. orthogonal transformations, linear transformations of this type are linear in all $k$-dimensional subspaces, so that the invariance of $T_{rs}$ follows from the fact that Hotelling's statistic is invariant under linear transformations in the $k$-dimensional subspaces.

### 3.3 Choosing $k$, $B_1$ and $B_2$

The choice of $k$ affects the performance of the $T_{rs}$ test. If $k$ is too close to $n_X+n_Y-2$ the test loses power, since Hotelling's $T^2$ performs poorly in this setting (Bai & Saranadasa, 1996). If $k$ is too small, much of the multivariate structure of the data set is lost.

Lopes et al. (2012) found analytically that the asymptotic power of their test was maximized when $k = \lfloor (n_X + n_Y - 2)/2 \rfloor$. We have verified numerically that this seems to be a good choice for $T_{rs}$ for finite sample sizes, although our simulation results indicate that the power is quite insensitive to small changes in $k$, as illustrated in Section 4.4. For large $n_X$ and $n_Y$, one may for purely computational reasons have to use a smaller $k$, as the permutation step may be computationally unfeasible when $k$ is too large.

The power of the test increases slightly with $B_1$ and $B_2$, but is relatively stable for $B_1, B_2 \geq 100$. In the simulation study below, we used $B_1 = 100$ and $B_2 = 500$.

## 4 Comparison of two-sample tests

### 4.1 Tests to be compared

In order to evaluate the performance of the random subspaces test and the random permutations version of the Lopes et al. (2012) test, we compared them to other two-sample tests.

Bai & Saranadasa (1996) proposed a Hotelling-type test utilizing the trace of the sample covariance matrix. More recently, Chen & Qin (2010) proposed a modification of the Bai–Saranadasa test, which has higher power in most situations. We therefore excluded the Bai–Saranadasa test from the study, choosing instead to focus our attention on the Chen–Qin test. We decided to use the asymptotic null distribution of the Chen–Qin statistic in our simulation, since using random permutations for computing the Chen–Qin p-value was extremely computer-intensive, at least using our implementation of the test.

Srivastava (2007) proposed using the Moore-Penrose inverse of the sample covariance matrix $\boldsymbol{S}$ when computing Hotelling's test statistic. In a small pilot study, we found the Srivastava test to be computationally expensive and to have lower power than the competing tests. It was therefore excluded from the larger study.

Srivastava & Du (2008) proposed replacing the sample covariance matrix $\boldsymbol{S}$ in Hotelling's statistic by a diagonal estimator. Their test statistic is asymptotically standard normal under certain conditions, but convergence to normality appears to be slow. We used random permutations to compute the p-values of the Srivastava–Du-test in our comparison. Compared to using the asymptotic null distribution, we found that the permutation procedure provided better type I error rates and resulted in higher power.

Finally, we considered test based on combined multiple $t$-tests. For such tests, the multivariate null hypothesis is rejected if at least one of the marginal null hypotheses is rejected. We used two methods for combining the tests: Bonferroni correction, controlling the family-wise error rate, and the Benjamini & Hochberg (1995) procedure, controlling the false discovery rate.

## 4.2 Invariance properties

When choosing between different tests, we are usually concerned with how well they attain their nominal type I error rates and how high their power is. In the high-dimensional setting, we must also take invariance properties into account, as discussed in Section 2.4. We assume that both samples are transformed analogously.

In Section 2.4 we argued that invariance under linear transformations of the marginal distributions is the most desirable invariance property in the gene expression setting. These are transformations of the type $\boldsymbol{X} \to \boldsymbol{D}\boldsymbol{X} + \boldsymbol{d}$, where $\boldsymbol{D}$ is a real diagonal $p \times p$ matrix with nonzero diagonal elements and $\boldsymbol{d} \in \mathbb{R}^p$. Among the tests considered here, only the marginal $t$-tests, the Srivastava–Du test and the random subspaces test (conditioned on the random subspaces chosen) are invariant under such transformations.

The Bai–Saranadasa, Chen–Qin and Srivastava tests are invariant under orthogonal transformations, i.e. transformations $\boldsymbol{X} \to c\boldsymbol{H}\boldsymbol{X}$, where $c$ is a nonzero constant and $\boldsymbol{H}$ is a real $p \times p$ orthogonal matrix. This invariance property is arguably less attractive in the gene expression setting, as rotations of the coordinate system are of little interest.

The Lopes et al. test is only invariant if the same scaling is applied to all marginal distributions, that is, under transformations of the type $\boldsymbol{X} \to c\boldsymbol{I}\boldsymbol{X}$, where $\boldsymbol{I}$ is the $p \times p$ identity matrix and $c$ is a non-zero constant.

## 4.3 Type I error rates

All tests under consideration are based on either asymptotic or random permutation estimates of the null distribution. They do therefore in general not attain the nominal type I error rates exactly. To evaluate the actual type I error rates for the tests in a few interesting settings, a Monte Carlo study was performed under the null hypothesis $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ for $p = 200$ and $n_X = n_Y = 50$.

Two families of distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ and three different covariance structures were evaluated in the simulations. Let $\boldsymbol{\Sigma}_{a,b}$ denote a covariance matrix with unit variances and 8 equal-sized blocks, where $\mathrm{Cov}(X_i, X_j)$ is $a$ if $X_i$ and $X_j$ belong to the same block and $b$ otherwise. The first three distributions were multivariate normal, with covariance matrices $\boldsymbol{\Sigma}_{0,0}$, $\boldsymbol{\Sigma}_{0.5,0.1}$ and $\boldsymbol{\Sigma}_{0.9,0.2}$. To study the impact of heavy-tailed distributions, the last two were multivariate $t$-distributions with 4 degrees of freedom and the covariance matrices $\boldsymbol{\Sigma}_{0,0}$ and $\boldsymbol{\Sigma}_{0.5,0.1}$.

For each distribution, $1,000$ samples were generated under the null hypothesis. The tests were then applied to each sample, using $k = 49$, $B_1 = 100$ and $B_2 = 500$ for $T_{rs}$ and $T_L$. The point estimates of the tests' type I error rates are given along with 95 % confidence intervals (not adjusted for multiplicity) in Table 1. All Hotelling-type tests have acceptable type I error rates, whereas the multiple $t$-tests have too low rates in some circumstances.

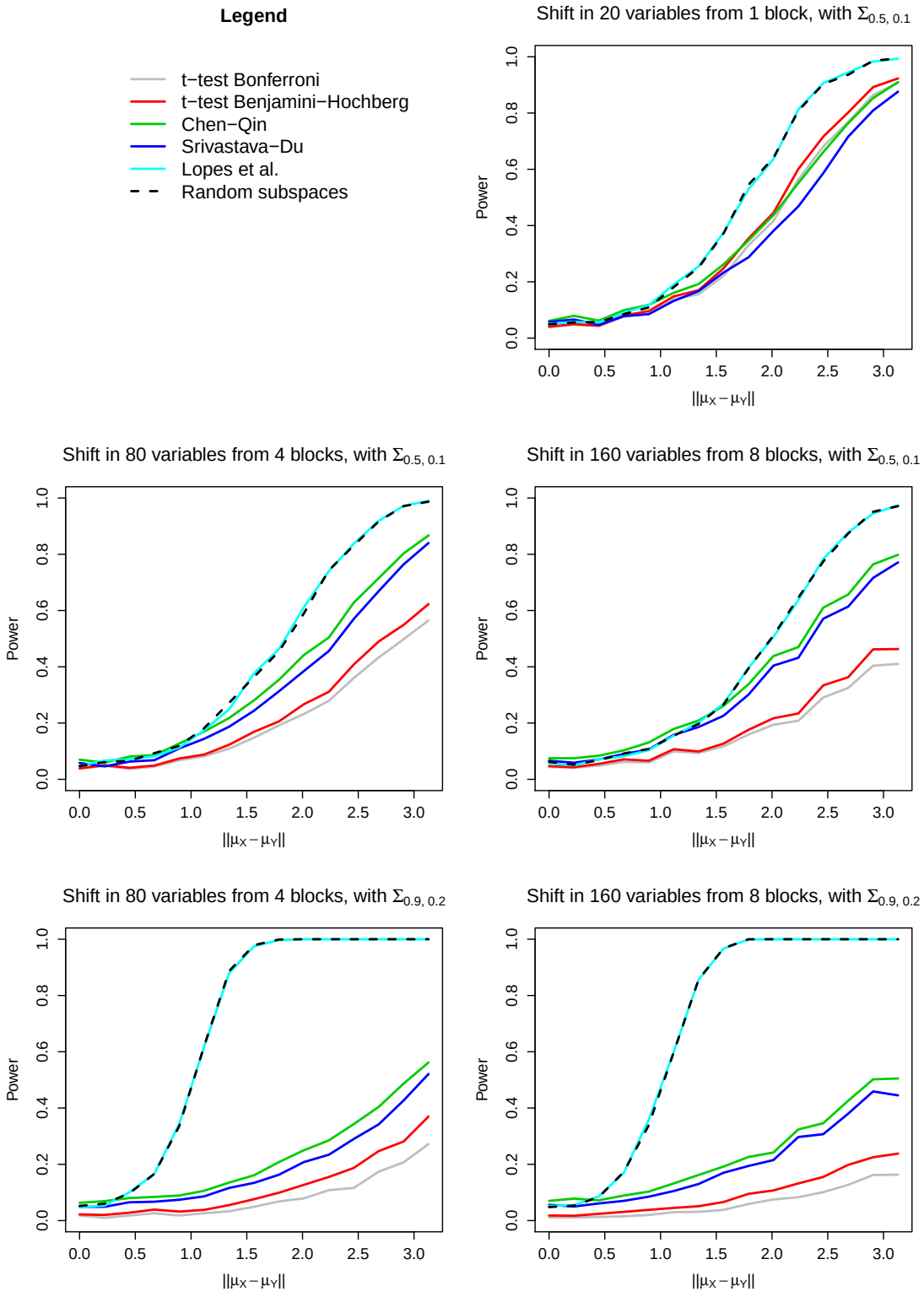
Table 1: Type I error rate of two-sample tests when $p = 200$, $n_X = n_Y = 50$ and $\alpha = 0.05$.

| | Bonferroni $t$ | Benjamini–Hochberg $t$ | Chen–Qin | Srivastava–Du | Lopes et al. $T_L$ | Random subspaces $T_{rs}$ |
|---|---|---|---|---|---|---|
| Normal $\boldsymbol{\Sigma}_{0,0}$ | 0.047 (0.035,0.061) | 0.048 (0.036,0.062) | 0.051 (0.039,0.066) | 0.046 (0.034,0.060) | 0.047 (0.035,0.061) | 0.049 (0.037, 0.063) |
| Normal $\boldsymbol{\Sigma}_{0.5,0.1}$ | 0.056 (0.043,0.072) | 0.058 (0.45,0.074) | 0.065 (0.051,0.081) | 0.052 (0.040, 0.067) | 0.046 (0.034,0.060) | 0.058 (0.045,0.074) |
| Normal $\boldsymbol{\Sigma}_{0.9,0.2}$ | 0.018 (0.011,0.028) | 0.022 (0.014,0.033) | 0.063 (0.049,0.079) | 0.049 (0.037,0.064) | 0.048 (0.036,0.0.63) | 0.052 (0.040,0.067) |
| $t(4)$ $\boldsymbol{\Sigma}_{0,0}$ | 0.028 (0.019,0.040) | 0.028 (0.019,0.040) | 0.056 (0.043,0.072) | 0.058 (0.045,0.074) | 0.046 (0.034,0.06) | 0.051 (0.039,0.066) |
| $t(4)$ $\boldsymbol{\Sigma}_{0.5,0.1}$ | 0.036 (0.026,0.049) | 0.039 (0.028,0.052) | 0.085 (0.069,0.103) | 0.071 (0.056,0.088) | 0.059 (0.046,0.075) | 0.066 (0.052,0.083) |

## 4.4 Power study

To study the power of the tests in the cancer pathway setting described in Section 1, we performed simulations under the assumptions of normality and the covariance matrices $\boldsymbol{\Sigma}_{0.5,0.1}$ and $\boldsymbol{\Sigma}_{0.9,0.2}$ described in the previous section. For the $\boldsymbol{Y}$ variable,

Figure 3: Power of two-sample tests when $p = 200$, $n_X = n_Y = 50$ and $\alpha = 0.05$.



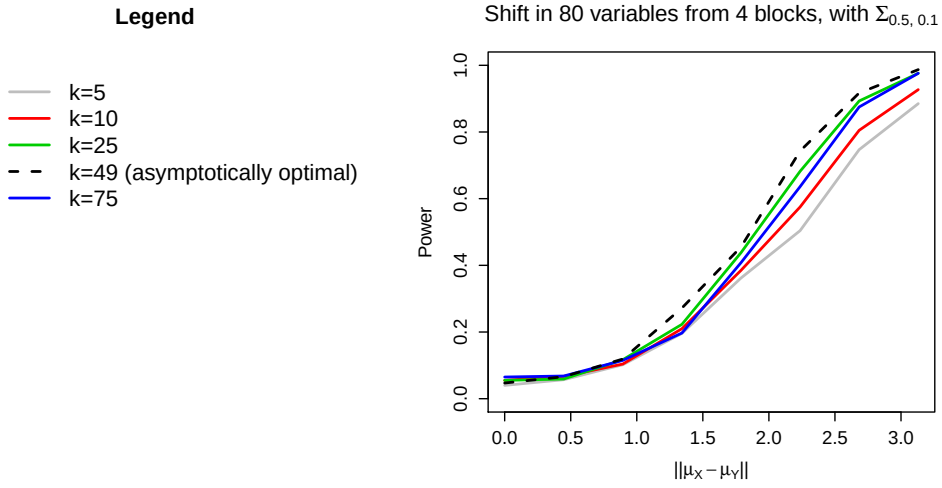
we shifted the means of 20 out of 25 genes evenly in each of $m$ out of the 8 blocks, for $m \in \{1, 4, 8\}$. The powers of the tests as functions of the resulting Euclidean distance $||\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y||$ are shown in Figure 3.

In the simulations, 1,000 samples were generated from each distribution. For the Lopes et al. and random subspaces tests, $B_1 = 100$, $B_2 = 500$ and $k = 49$ were used. We also performed comparisons for other $p$, $n_X$ and $n_Y$, as well as settings where $\boldsymbol{\Sigma}_X \neq \boldsymbol{\Sigma}_Y$ and where the shifts were unevenly distributed among the variables. The resulting plots were not qualitatively different from those in Figure 3 and are therefore not shown here.

In our comparison, the Lopes et al. test and the random subspaces tests were the only tests that *gained* power as the correlation between the variables was increased. For all the other tests, the power became lower when the variables became more dependent. For a fixed covariance matrix, the power of the Hotelling-type tests are relatively stable when the number of variables in which there is a difference change. The multiple $t$-tests are much more sensitive to this type of changes: their power decreases as the number of shifted variables increases.

In Figure 4 we plot the power of $T_{rs}$ for different choices of $k$ for one of the alternatives from Figure 3. The plots for other alternatives are similar. While $k = \lfloor (n_X + n_Y - 2)/2 \rfloor$ gives the highest power, the test is surprisingly insensitive to changes in $k$. In the $p = 200$, $n_X = n_Y = 50$ setting, where $k = 49$ is asymptotically optimal, there is little difference in power when $25 \leq k \leq 75$. Under this particular alternative, the test is on a par with the Chen–Qin and Srivastava–Du tests even when $k = 5$ (cf. Figure 3).

Figure 4: Power of the random subspaces test for different $k$ when $p = 200$, $n_X = n_Y = 50$ and $\alpha = 0.05$.



# 5 Computational aspects

The random subspaces test quickly becomes very computer-intensive as $k$ increases, as we must invert $B_1 \cdot B_2$ matrices of size $k \times k$ in order to obtain the p-value. In this section we discuss some ways to speed up the computations and compare the performances of different implementations in R.

Small gains in speed can be achieved by avoiding repeating the same calculations more than once, e.g. by computing $n = n_X + n_Y - 2$ only once. Moreover, when using the permutation distribution instead of the asymptotic null distribution, there is for instance no need to rescale the test statistic by the sample size.

The most important tool for improving the computational speed, however, is parallelization, i.e. running the repetitions of steps 1-2 of Algorithm 2 simultaneously instead of sequentially. An efficient parallel R implementation of the random subspaces test is given in Appendix A.

The choice of $k$ impacts not only the power of the test, as shown in Figure 4, but also the computational cost of the algorithm. There is therefore a trade-off between computational speed and power; $k = \lfloor (n_X + n_Y - 2)/2 \rfloor$ yields the highest power, but smaller $k$ will give improved computational performance. Similarly, larger $B_1$ and $B_2$ are preferable from a statistical viewpoint, but lead to increased computational costs.

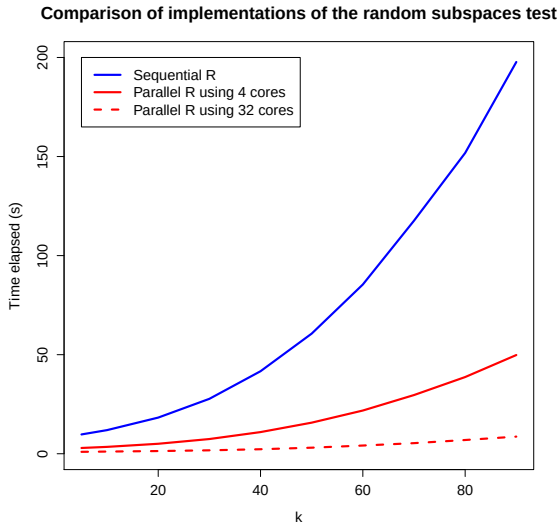
In Figure 5 the performances of two implementations of the test are compared for different choices of $k$ when $p = 200$ and $n_X = n_Y = 50$. The first is a naive sequential implementation in R. The second is a parallel implementation in R, using the `doMC` package. The implementations were executed on a 64-core 2.2 GHz AMD processor with 128 GB RAM and R 2.15.2, using different numbers of cores. The difference between the sequential and parallel R implementations is quite striking: when $k = 50$, the sequential implementation needed 61 second to perform the test, whereas the parallel implementation only needed 15.7 seconds using 4 cores and 3.0 seconds using 32 cores.

# 6 Discussion

## 6.1 Recommendations for two-sample tests

Non-negligible dependence structures are present in virtually all genetic datasets and tests for differentially expressed gene-sets should therefore take such dependencies into account. Most high-dimensional two-sample tests, as well as multiple $t$-tests, do not do this to a sufficient degree. The Lopes et al. and random subspaces tests do account for dependencies, and consequently outperform their competitors in settings with non-negligible dependence structures. The powers of the two tests

Figure 5: Mean execution time of a single random subspaces test for different $k$ when $p = 200$ and $n_X = n_Y = 50$.




closely mimic each other. Unlike the Lopes et al. test however, the random subspaces test has the additional benefit of invariance under linear transformations of the marginals. Judging from the simulation results in Section 4, we therefore recommend the random subspaces test as the default high-dimensional two-sample test.

## 6.2 Testing multiple gene-sets

In most studies, one wishes to perform tests for a large number of gene-sets and not just for a single set. It is generally agreed upon that one should use some sort of adjustment for multiplicity in such investigations, in order to lower the number of false discoveries. Methods for doing such adjustments include Bonferroni corrections for controlling the family-wise error rate (e.g. Holm, 1979), the Benjamini & Hochberg (1995) false discovery rate procedure and resampling methods (Dudoit et al., 2003; Ge et al., 2003; Barry et al., 2005; Dudoit & van der Laan, 2008; Sohn et al., 2011). The latter class of methods seems especially promising when dependent gene-sets are tested, as resampling methods to a greater extent can account for such dependencies. When used with test procedures that use random permutations to compute p-values, such methods can however become extremely computer-intensive. It should also be noted that research on multiple testing in genetic research mostly has focused on gene-level $t$-tests and that recommenda-

tions for univariate test need not carry over to the multivariate setting. How to adjust high-dimensional tests for multiplicity remains an open problem.

## Acknowledgments

The author wishes to thank Elisabeth Thulin and Silvelyn Zwanzig for helpful discussions. Miles Lopes kindly provided R code for computing the Lopes et al. statistic, allowing the author to compare their implementation of the test to his own.

# References

Bai, Z., Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, **6**, 311–329.

Barry, W., Nobel, A., Wright, F. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.

Barry, W., Nobel, A., Wright, F. (2008). A statistical framework for testing functional categories in microarray data. *Annals of Applied Statistics*, **2**, 286–315.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289–300.

Bertoni, A., Folgiere, R., Valentini, G. (2005). Bio-molecular cancer prediction with random subspace ensembles of support vector machines. *Neurocomputing*, **63**, 535–539.

Bingham, E., Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data, in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'01)*.

Chen, S.X., Qin, Y.L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics*, **38**, 808–835.

Cuesta-Albertos, J.A., Fraiman, R., Ransford, T. (2006). Random projections and goodness-of-fit tests in infinite-dimensional spaces. *Bulletin of the Brazilian Mathematical Society*, **37**, 477–501.

Dudoit, S., Shaffer, J.P., Boldrick, J.C. (2008). Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71–103.

Dudoit, S., van der Laan, M.J. (2008). *Multiple testing procedures with applications to genomics*, Springer, New York.

Durand, J., Atkison, T. (2011). Using randomized projection techniques to aid in detecting high-dimensional malicious applications, in *Proceedings of the 49th Annual Southeast Regional Conference (ACM-SE'11)*.

Efron, B., Tibshirani, R., Storey, J., Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151–1160.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, **102**, 93–103.

Efron, B., Tibshirani, R. (2007). On testing the significance of sets of genes. *Annals of Applied Statistics*, **1**, 107–129.

Fern, X.Z, Brodley, C.E. (2003). Random projection for high dimensional data clustering: a cluster ensemble approach, in *Proceedings of 20th International Conference on Machine learning (ICML2003)*.

Gatti, D.M., Barry, W.T., Nobel, A.B., Rusyn, I., Wright, F.A. (2010). Heading down the wrong pathway: one the influence of correlation within gene sets. *BMC Genomics*, **11**, 574.

Ge, Y., Dudoit, S., Speed, T. (2003). Resampling-based multiple yesting for microarray data analysis. *Test*, **12**, 1–77.

Goeman, J.J., Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Henrion, M., Mortlock, D.J., Hand, D.J., Gandy, A. (2011). Subspace methods for anomaly detection in high dimensional astronomical databases, in *Proceedings of the 58th World Statistics Congress of the International Statistical Institute (ISI11)*.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Jacob, L., Neuival, P., Dudiot, S. (2012). More power via graph-structured tests for differential expression of gene networks. *Annals of Applied Statistics*, **6**, 561–600.

Lai, C., Reinders, M.J.T., Wessels, L. (2006). Random subspace method for multivariate feature selection. *Pattern Recognition Letters*, **27**, 1067–1076.

Lehmann, E.L. (1959). *Testing Statistical Hypotheses*, John Wiley & Sons, New York.

Lopes, M.E., Jacob, L.J., Wainwright, M.J. (2012). A more powerful two-sample test in high dimensions using random projection, arXiv:1108.2401v2.

Nettleton, D., Recknor, J., Reecy, J. (2008). Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics*, **24**, 192–201.

Newton, M., Quintana, F., Den Boon, J., Sengupta, S., Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Annals of Applied Statistics*, **1**, 85–106.

Qui, X., Klebanov, L., Yakovlev, A. (2005). Correlation between gene expression levels and limitations of the empirical Bayes methodology in microarray data analysis. *Statistical Applications in Genetics and Molecular Biology*, **4**, 34.

Sohn, I., Owzar, K., Lim, J., George, S.L., Mackey Cushman, S., Jung, S.-H. (2011). Multiple testing for gene sets from microarray experiments. *BMC Bioinformatics*, **12**, 209.

Srivastava, M.S. (2007). Multivariate theory for analyzing high dimensional data. *Journal of The Japan Statistical Society*, **37**, 53–86.

Srivastava, M.S., Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, **99**, 386–402.

Srivastava, M.S., Katayarna, S., Kano, Y. (2013). A two sample test in high dimensional data. *Journal of Multivariate Analysis*, **114**, 349–358.

Varmuza, K., Filzmoser, P., Liebmann, B. (2010). Random projection experiments with chemometric data. *Journal of Chemometrics*, **24**, 209–217.

Vempala, S.S. (2004). *The Random Projection Method*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, ISBN 0-8218-3793-1.

Wei, S., Lee, C., Wichers, L., Li, G., Marron, J.S. (2013). Direction-projection-permutation for high dimensional hypothesis tests, arXiv:1304.0796v1.

Contact information:
Måns Thulin, Department of Mathematics, Uppsala University, Box 480, 751 06 Uppsala, Sweden
E-mail: thulin@math.uu.se

# A Appendix: An R implementation of the random subspaces test

An implementation of the random subspaces statistic in R is given below. In order to speed up the computation of the statistic, the `.Internal` versions of `mean` and `cov` are used. This means that the function avoids error handling, such as checking whether the data contains `NA` values. The outer loop used in the computation is parallelized using the `foreach` and `doMC` packages.

```
# Load required packages:
library(compiler) # Compilation - for better performance
library(doMC)     # Parallelization - for better performance
registerDoMC()

# Hotellings T^2 statistic
T2.func<-function(x,y,n1,n2,p)
{
    mdiff<-.Internal(colMeans(x,n1,p,na.rm=FALSE))-
           .Internal(colMeans(y,n2,p,na.rm=FALSE))
    Spool<-((n1-1)*.Internal(cov(x, NULL, 1, FALSE))+(n2-1)*
                  .Internal(cov(y, NULL, 1, FALSE)))/(n1+n2-2)
    return(t(mdiff) %*% solve((1/n1+1/n2)*Spool) %*% mdiff)
}
T2.func<-cmpfun(T2.func)

# Random subspaces
subspacesT2<-function(x,y,n1,n2,p,B=100,k=floor((n1+n2-2)/2))
{
    res<-vector(length=B)
    for(j in 1:B)
    {
      x.cols<-sample(p,k)
      x.new<-x[,x.cols]
      y.new<-y[,x.cols]
      res[j]<-T2.func(x.new,y.new,n1,n2,k)
   }
   return(.Internal(mean(res)))
}
subspacesT2<-cmpfun(subspacesT2)

subspaces.test<-function(x,y,n1,n2,p,B1=100,B2=100,k=floor((n1+n2-2)/2))
```

```
{
    z<-rbind(x,y) # Big matrix to resample from
    zsize<-n1+n2

    # Permutations
    rs<-data.frame(foreach(i = 1:B1) %dopar%
    {
      x.rows<-sample(zsize,n1)
      x.new<-z[x.rows,]
      y.new<-z[-x.rows,]
      subspacesT2(x.new,y.new,n1,n2,p,B2,k)
    })

    rs.obs<-subspacesT2(x,y,n1,n2,p)

    # Return p-value:
    return(sum(rs>=as.numeric(rs.obs))/B1)
}
subspaces.test<-cmpfun(subspaces.test)

# Example of usage:
library(MASS)    # Used to generate multivariate normal data

# Set parameters:
p<-200
n1<-n2<-50
mu1<-rep(1,p)
mu2<-rep(1.02,p)
Sigma1<-matrix(0.25,p,p)
diag(Sigma1)<-1

# Generate example data:
x<-mvrnorm(n1, mu1, Sigma1)
y<-mvrnorm(n2, mu2, Sigma1)

# Apply test:
subspaces.test(x,y,n1,n2,p,B1=100,B2=500,k=49)
```

# B Appendix: Invariance under linear transformations of the marginals

The example below illustrates the lack of invariance of the Lopes et al. test, discussed in Section 2.4. It uses the function `T2.func` from the previous section.

```
library(doMC)
registerDoMC()
library(MASS)
library(compiler)

# Compute the random projections statistic
projectionT2<-function(x,y,n1,n2,p,B=100,k=floor((n1+n2-2)/2),matrixlist)
{
   res<-vector(length=B)
   for(j in 1:B)
   {
      P<-as.matrix(data.frame(matrixlist[j]))
      x.new<-x %*% P
      y.new<-y %*% P
      res[j]<-T2.func(x.new,y.new,n1,n2,k)
   }
   return(.Internal(mean(res)))
}
projectionT2<-cmpfun(projectionT2)

# Set parameters:
p<-20
n1<-5
n2<-5
k<-4
B<-100
B2<-500

# Import data:
x<-matrix(c(1.46,-2.28,0.73,0.02,0.39,0.75,-0.43,1.9,1.23,-0.15,1.31,2.37,
5.37,-2.03,-1.48,5.16,-2.72,-2.67,3.9,-0.97,0.84,-2.55,1.37,-1.53,1.7,
4.03,-0.1,0.97,4.24,0.43,3.13,-5.38,0.13,4.67,6.01,2.75,-1.71,3.52,2.17,
-2.93,2.45,2.59,-1.59,5.64,4.8,5.01,3.15,4.36,5.27,-0.53,1.58,0.53,1.39,
1.67,0.16,1.32,0.61,1.54,1.81,0.59,1.4,2.17,1.8,0.34,0.74,0.06,1.24,1.44,
0.91,0.55,0.22,1.32,0.36,0.94,1.34,1.87,0.69,0.65,1.62,0.16,0.28,-0.3,0.84,
```

```
1.28,1.33,2.3,1.55,1.54,1.87,1.29,2.2,0.9,1.44,2.02,1.34,1.73,1.92,0.31,
0.81,0.75),5,20)

y<-matrix(c(1.57,2.03,0.58,3.2,2.03,4.18,3.04,0.4,1.73,2.3,3.2,2.84,3.39,
4.17,2.84,1.26,2.88,-1.07,4.4,-0.71,-5.21,-2.07,2.7,6.02,-1.38,-0.03,3.06,
0.29,4.15,2.02,3.07,3.86,5.81,2.62,0,4.66,3.3,0.37,2.57,4.57,3.86,3.46,-1,
2.72,-1.58,2.06,6.09,6.88,1.36,0.07,2.33,-0.17,2.37,1.85,1.15,3.77,1.1,
2.15,2.5,1.52,3.31,-0.2,2.7,1.89,1.8,2.61,1.34,2.55,3.87,1.58,3.29,2.97,
2.33,2.6,2.65,2.32,1.23,2.22,2.83,1.52,1.85,1.84,1.8,1.75,1.96,1.31,0.47,
1.3,2.69,2.01,3.41,1.03,1.44,0.65,1.76,1.72,0.85,3.58,1.2,2.2),5,20)

# Standardize data:
S<-cov(rbind(x,y))
C<-diag(1/sqrt(diag(S)))
x2<-x %*% C
y2<-y %*% C

# Decide which matrices to use for the random projections:
matrixlist<-foreach(j = 1:B) %dopar%
{
   matrix(rnorm(k*p),p,k)
}

# Compute test statistics:
T.raw<-projectionT2(x,y,n1,n2,p,B,k,matrixlist)
T.standardized<-projectionT2(x2,y2,n1,n2,p,B,k,matrixlist)

# Permutations:
z<-rbind(x,y)
z2<-rbind(x2,y2)
zsize<-n1+n2

rs<-data.frame(foreach(i = 1:B2) %dopar%
{
   matrixlist2<-foreach(j = 1:B) %dopar%
   {
      matrix(rnorm(k*p),p,k)
   }
   x.rows<-sample(zsize,n1)
   x.new<-z[x.rows,]
```

```
   y.new<-z[-x.rows,]
   x2.new<-z2[x.rows,]
   y2.new<-z2[-x.rows,]
   c(raw=projectionT2(x.new,y.new,n1,n2,p,B,k,matrixlist2),
   stand=projectionT2(x2.new,y2.new,n1,n2,p,B,k,matrixlist2))
})

raw<-as.numeric(rs[1,])
stand<-as.numeric(rs[2,])

# p-value for raw data:
sum(raw>=as.numeric(T.raw))/B2

# p-value for standardized data:
sum(stand>=as.numeric(T.standardized))/B2
```